УДК 519.2/.6

МЕТОД ГЛАВНЫХ КОМПОНЕНТ И ЛИНЕЙНЫЕ МНОГООБРАЗИЯ НА ПРАКТИКЕ: ПРИМЕНЕНИЕ К РОССИЙСКОЙ БАНКОВСКОЙ СИСТЕМЕ

Веретнова К.Ю.
научный руководитель канд. техн. наук Покидышева Л. И.
Институт математики,
Сибирский Федеральный Университет

Существует множество агентств по созданию рейтингов. Количество подходов к созданию рейтинга между объектами очень велико. Создаются формулы, по которым тот или иной объект исследования ставится на то или иное место в рейтинге. В данной работе предложена альтернатива «искусственным», придуманным рейтингам. Проводится работа по созданию «объективного» рейтинга среди банков, действующих на территории Российской Федерации, не зависящего от выбора показателей банков, от какой-либо не объективной формулы. С применением нелинейного моделирования, использующего метод главных компонент и главные многообразия был составлен рейтинг банков, действующих на территории Российской Федерации.

Есть несколько подходов к созданию рейтинга между какими-либо объектами. Объекты каким-то образом располагаются в пространстве своих показателей, которыми они характеризуются. Один из подходов – это создание некой прямой шкалы рейтинга, на которую проецируются объекты в пространстве показателей. Этот подход не является «естественным». Александром Николаевичем Горбанем и Андреем Юрьевичем Зиновьевым была предложена идея другого подхода к созданию рейтингов, которая применяется в тех случаях, когда линейная аппроксимация данных не является удовлетворительной. Если объекты располагаются в пространстве вдоль некоторой кривой, то более естественный способ расположить объекты в рейтинге - аппроксимировать облако объектов в пространстве показателей некоторой гладкой кривой и проецировать объекты на нее, а не на придуманную шкалу. Эта линия – натуральная шкала рейтинга. Она будет учитывать нелинейные структуры облака данных. Такую линию мы будем называть главным многообразием.

Данный подход позволяет не изобретать формул для определения позиции объекта в рейтинге, а так же позволяет отказаться от выбора показателей. Данная кривая появляется как аппроксимация облака точек. Место объекта в рейтинге определяется координатой на этой «натуральной» кривой.

Цель работы: проверить действительно ли большой набор данных может быть аппроксимирован линейным многообразием меньшей размерности. И в рамках примера: создать рейтинг между банками, используя метод упругих карт.

Метод главных компонент - это один из способов понижения размерности данных, состоящий в переходе к новому ортогональному базису, оси которого ориентированы по направлениям максимальной дисперсии набора входных данных.

После того как два вектора главных компонент будут найдены, точки пространства показателей (банки) можно спроецировать на плоскость, образованную векторами главных компонент.

Для применения метода главных компонент данные должны быть записаны в виде матрицы. Отдельная строка такой матрицы – это конкретный объект исследования; вектор, с координатами-показателями.

Зачастую объекты исследования имеют сильно различающиеся значения показателей по тому или иному показателю. Для того чтобы уравновесить эти

значения, данные необходимо нормировать по столбцам. Так же метод главных компонент предполагает, что вектора данных являются центрированными.

Таким образом, вектор-столбец P_j j-го показателя следует нормировать по следующей формуле:

$$P_j = \left(\frac{P_j^1 - \overline{P_j}}{\sigma(P_j)}, \cdots, \frac{P_j^n - \overline{P_j}}{\sigma(P_j)}\right),\ j = \overline{1,m}$$

где, матрица данных имеет размерность $(n \times m)$, $\overline{P_j}$ – выборочное среднее для j-го показателя P_j,

$\overline{P_j} = \frac{1}{n}\sum_{i=1}^{n} P_j^i$, $\sigma(P_j)$ – среднеквадратическое отклонение для j-го показателя P_j, $\sigma(P_j) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_j^i - \overline{P_j})}$.

Вектора главных компонент для показателей были найдены как собственные вектора эмпирической ковариационной матрицы. Благодаря нормировке ковариационная матрица стала корреляционной. Вектора и собственные значения матрицы были найдены с помощью метода вращений Якоби.

Упругая карта служит для нелинейного сокращения размерности данных. В многомерном пространстве данных располагается поверхность, которая приближает имеющиеся точки данных и при этом является, по возможности, не слишком изогнутой. Данные проецируются на эту поверхность и потом могут отображаться на ней, как на карте. Ее можно представлять себе как упругую пластину, погруженную в пространство данных и прикрепленную к точкам данных пружинками. Служит обобщением метода главных компонент (в котором вместо упругой пластины используется абсолютно жесткая плоскость).

По построению, упругая карта представляет собой систему упругих пружин, вложенную в многомерное пространство данных. Эта система аппроксимирует облако данных. Метод был разработан проф., д.ф.-м.н. А. Н. Горбанем, к.т.н. А. Зиновьевым и к.т.н. А. Питенко в 1996—2001 гг.

Рассмотрим двумерную прямоугольную сетку узлов, в которой p узлов по горизонтали и q узлов по вертикали. Узлы сетки нумеруются с помощью двух индексов $y^{ij}, i = \overline{1,p}, j = \overline{1,q}$.

Определение 1. *Упругой сеткой* будем называть множество узлов таких что:

1) Узлы сетки близки к точкам данных;
2) Сетка должна быть упруга по отношению к изгибу. Это свойство до некоторой степени обеспечит гладкость результирующего многообразия;
3) Сетка должна быть равномерна, то есть упруга по отношению к растяжению.

Меняя параметры упругости и растяжения можно получать сетку с различными свойствами.

Определение 2. *Таксон K_{ij} узла y^{ij}* – это множество точек $x \in X$, таких что:

$$K_{ij} = \left\{x \in X \mid \left\|y^{ij} - x\right\|^2 \to min\right\},$$

где X – множество точек данных, $\|\cdot\|$ – Евклидова норма.

Функционал степени качества сетки, который необходимо минимизировать:

$$D = \frac{D_1}{|X|} + \lambda\frac{D_2}{pq} + \mu\frac{D_3}{pq} \to min,$$

$|X|$ – число точек данных, λ, μ - коэффициенты упругости, отвечающие за растяжение и изогнутость сетки.

$D_1 = \sum_{ij}\sum_{x_k \in K_{ij}}\left\|x_k - y^{ij}\right\|^2$ – мера близости узлов к данным,

$D_2 = \sum_{i=1}^{p}\sum_{j=1}^{q-1}\left\|y^{ij} - y^{i,j+1}\right\|^2 + \sum_{i=1}^{p-1}\sum_{j=1}^{q}\left\|y^{ij} - y^{i+1,j}\right\|^2$ – мера растянутости сетки,
$D_3 = \sum_{i=1}^{p}\sum_{j=2}^{q-1}\left\|2y^{ij} - y^{i,j-1} - y^{i,j+1}\right\|^2 + \sum_{i=2}^{p-1}\sum_{j=1}^{q}\left\|2y^{ij} - y^{i-1,j} - y^{i+1,j}\right\|^2$ – мера кривизны сетки.

Так как метрика является евклидовой, то функционал D является квадратичным по отношению к узлам y^{ij}. Следовательно, для его минимизации будет приемлем следующий алгоритм:

Шаг 1. Узлы сетки, так или иначе, располагаются в пространстве.

Шаг 2. При заданных положениях узлов данные разбиваются на таксоны.

Шаг 3. При заданном разбиении множества точек данных на таксоны производится минимизация функционала D из условия $\frac{\partial D}{\partial y^{ij}} = 0$.

Шаги 2 и 3 повторяются до тех пор, пока величина функционала D не станет мала (в пределах заданной точности).

В настоящей работе были исследованы первые 100 банков, действующих на территории Российской Федерации, из рейтинга, опубликованного на сайте Центрального банка РФ. Каждый банк характеризуется 74 значениями показателей. Показатели включают в себя кредиты, выданные как физическим лицам, так и различным коммерческим организациям, на различные периоды, депозиты коммерческих, некоммерческих организаций, физических лиц, основные средства и т.п.

Таким образом, мы можем представить множество исследуемых банков как облако в 74-размерном пространстве показателей.

С помощью метода упругих карт была получена аппроксимация облака банков кривой, на которую банки были спроецированы.

После того как вектора главных компонент были найдены, данные были спроецированы на плоскость, образованную первыми двумя вектора главных компонент.

После операции проецирования был получен рисунок облака данных на плоскости главных компонент (Рис. 1).

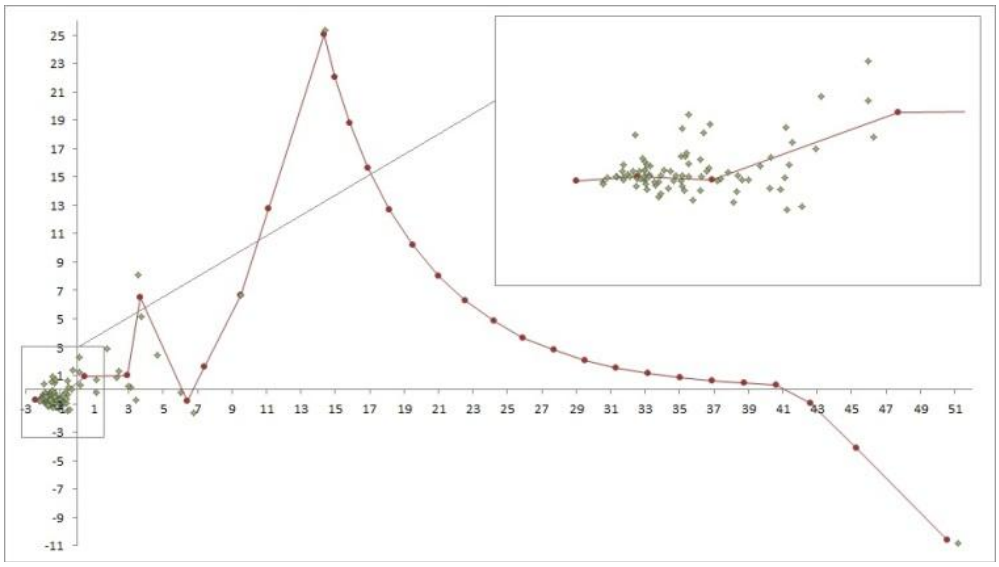

Рис.1: Проекция точек данных на плоскость главных компонент. Так же на рисунке представлена кривая, аппроксимирующая облако данных.

После того как многообразие построено, для визуализации данных необходимо указать правило, с помощью которого данные из исходного пространства переносятся

на упругую кривую. Длина вектора переноса не будет слишком велика, поскольку карта аппроксимирует данные и достаточно плотно к ним прилегает.

Идея, которая применяется при проецировании – сопоставление точке данных ближайшей точки отрезка, соединяющего два ближайших к точке узла на карте. Место объекта в рейтинге определятся координатой точки на кривой, аппроксимирующей облако данных.

Определение 3. *Расстояние* будем определять следующим образом: выполним ортогональное проецирование на прямую, содержащую отрезок. Если проекция принадлежит отрезку, то искомое расстояние – это расстояние до проекции. Иначе искомое расстояние – это расстояние до ближайшего конца отрезка. Тогда координатой точки на кривой будет либо координата этой точки на отрезке, либо координата ближайшего к ней узла.

Так как отрезок является линейной комбинацией координат двух точек, тогда условием принадлежности точки отрезку будет существование такого $p \in [0,1]$ что:

$$\begin{cases} px_1 + (1-p)x_2 = x \\ py_1 + (1-p)y_2 = y \end{cases}$$

где (x, y)– координаты точки данных, (x_1, y_1), (x_1, x_2) – координаты ближайших к ней узлов.

Рассмотрим Рис.2. Точки на кривой – проекции точек-банков на линейное многообразие. Все банки выстроились один за другим. Выпрямляя кривую, мы получим «естественный» рейтинг банков.

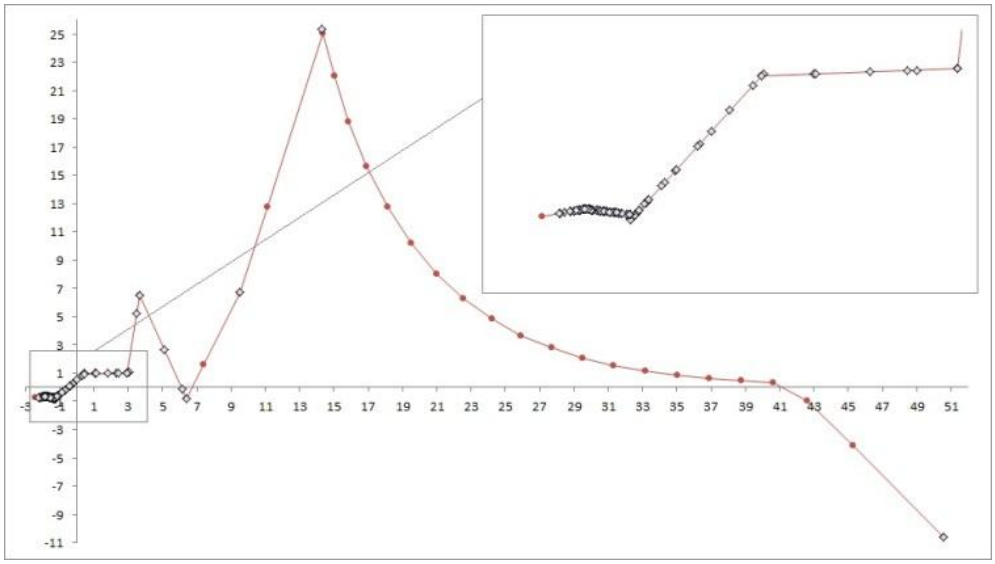

Рис.2: Проекции точек данных на кривой.

Таким образом, применяя метод главных компонент и метод упругих карт можно получать рейтинг, независимый от чьей-либо точки зрения. Аппроксимируя облако данных кривой, мы получаем «натуральную» шкалу рейтинга, которой можно пользоваться.